Show and tell: learning causal structures from observations and explanations

Andrew Nam, Christopher Hughes, Thomas Icard, Tobias Gerstenberg
Stanford University

Introduction

Prior work has looked at how people infer causal structures from observational data and interventions. Here, we look at what people can learn from explanations.

Pearl's causal hierarchy describes three levels of causal reasoning

Level	Name	Symbol	Activity	Example
1	Association	P(y|x)	Observing	What does a symptom tell me about a disease?
2	Intervention	P(y|do(x), z)	Intervening	If I take aspirin, will my headache be cured?
3	Counterfactuals	$P(y_x\|x', y')$	Imagining, retrospecting	Was it the aspirin that stopped my headache?

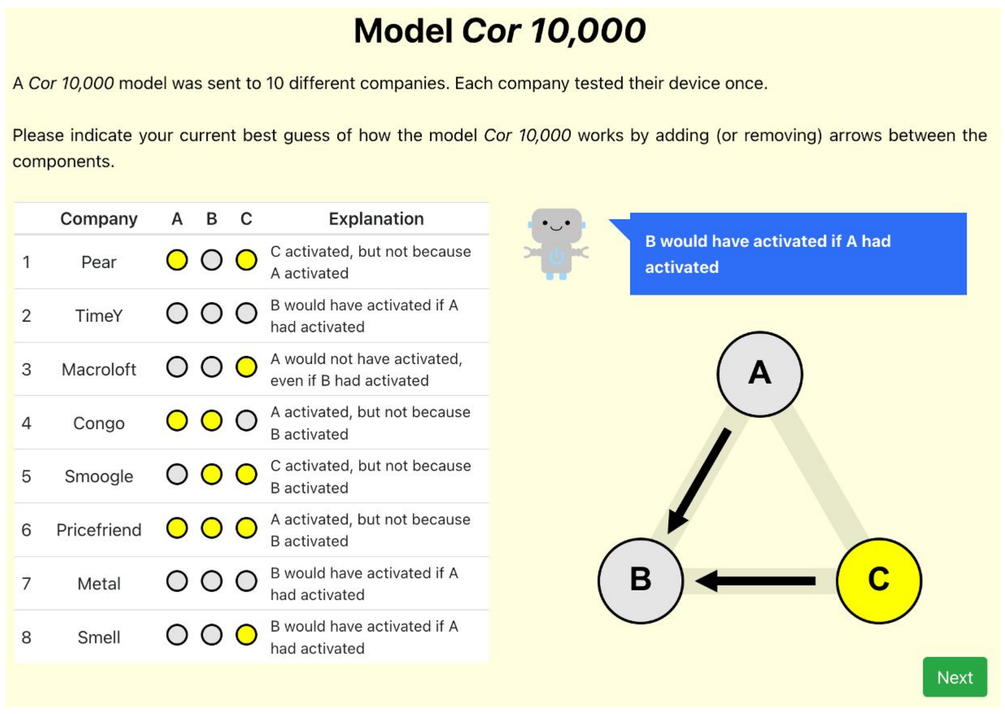
Explanations provide information at the third level by describing counterfactual dependencies.

Factual (explanation): The aspirin stopped the headache. ⟷ **Counterfactual**: If I hadn't taken the aspirin, I would still have the headache.

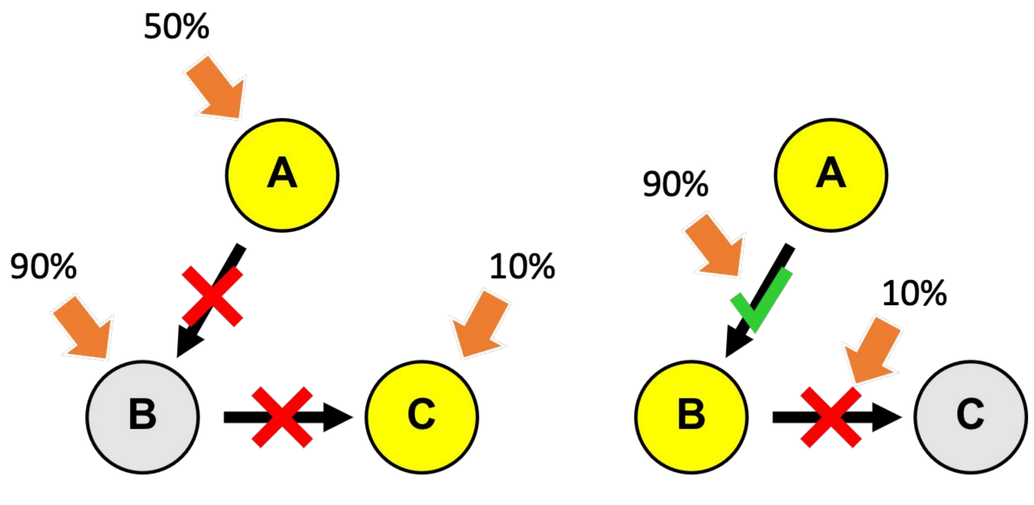
Causal Inference Task

Goal: Infer the connectivity of a probabilistic device (a 3-node directed acyclic graph) from observational and/or explanatory data across 10 trials.

- On each trial, participants are shown the node activations (observational data) and/or a description (explanatory data) about a connection from an event (which nodes and connections activated)
- Using the cumulative information shown across trials, participants indicate what they infer to be the true connectivity of the device

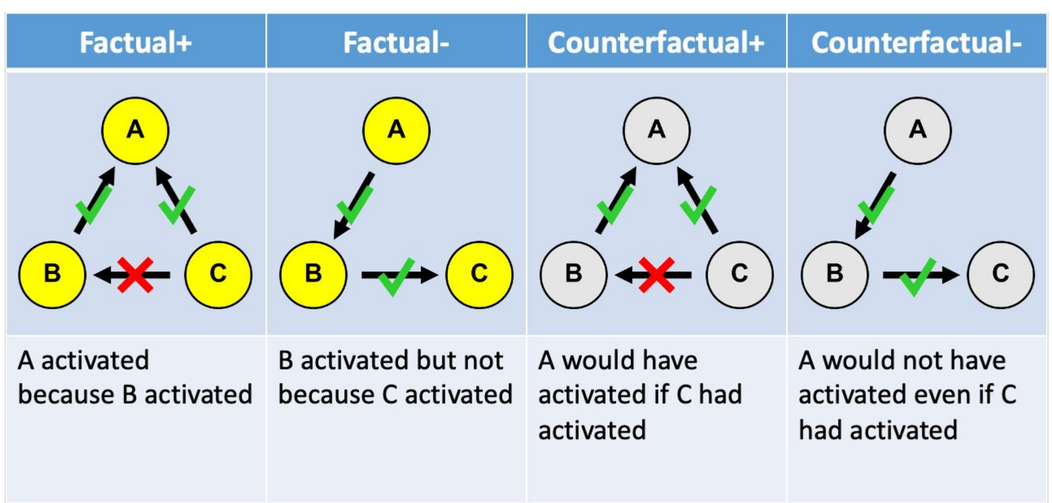


Device properties



Explanation properties

- Describes the actual event truthfully
- Describes two nodes with the same activation states
- May describe an indirect causation
- A valid explanation is sampled randomly to be shown to the participant (not guaranteed to be helpful)



Computational Model

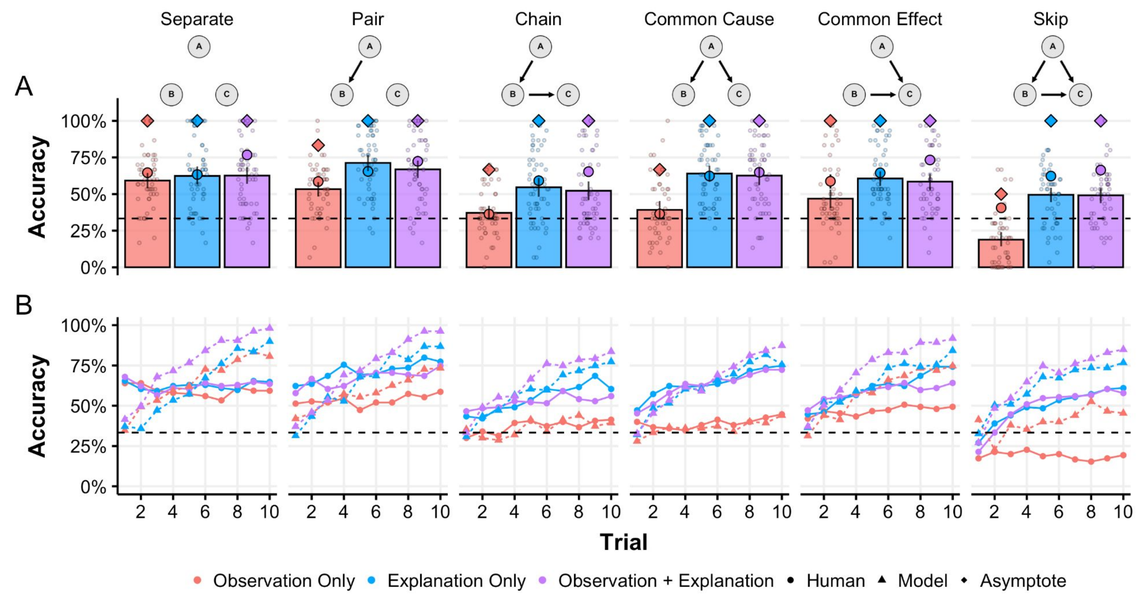
Using Bayes' theorem, we compute the posterior probability of each device d_i given the data

$$P^{(t)}(d_i|o^{(t)},x^{(t)}) = \frac{\sum_{e\in \mathrm{E}} \mathbb{1}(o^{(t)} \text{ in } e)P(x^{(t)}|e)P(e|d_i)P^{(t-1)}(d_i)}{\sum_{d\in \mathrm{D}}\sum_{e\in \mathrm{E}} \mathbb{1}(o^{(t)} \text{ in } e)P(x^{(t)}|e)P(e|d)P^{(t-1)}(d)}$$

- $P^{(t)}(d_i|o^{(t)},x^{(t)})$: The probability of device d_i at trial t given the observation o and explanation x
- $P(e|d_i)$: The probability of event e (set of active nodes and working connections) under device d_i
- $P(x^{(t)}|e)$: The probability of the shown explanation under event e
- $\mathbb{1}(o^{(t)} \text{ in } e)$: Whether the observed activations are consistent with the event

For models of observation-only and explanation-only conditions, we remove all terms containing o or x, respectively.

Analysis 1: Overall Accuracy



Number of correctly inferred edges averaged all trials (A) and on individual trials (B). Bars in (A) indicate human accuracy, circles indicate model accuracy, and diamonds indicate highest possible accuracy with infinite trials.

1. People generally are significantly suboptimal in their inferences compared to Bayes optimality.
2. Explanations provide significantly more helpful signals than observations alone, and usefulness increases with number of edges in the graph.
3. Having both observations and explanations show no significant improvement over having only explanations.

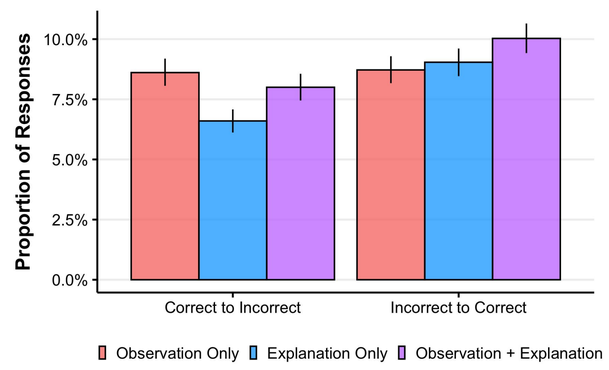
Analysis 2: Differences in Response Types

There are two ways to improve accuracy

1. Increase *incorrect* → *correct* revisions
2. Decrease *correct* → *incorrect* revisions

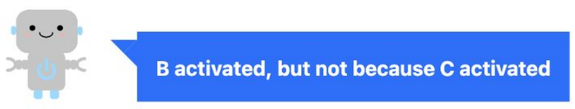
We find that

1. Explanations help reduce revising already correct inferences.
2. Explanations have little advantage in making new correct inferences.
3. Observations increase the number of revisions when given alongside explanations.

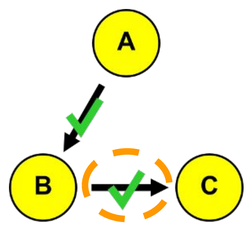


Analysis 3: Interpreting Negative Explanations

Negative explanations are less common and more ambiguous. How would people use them?



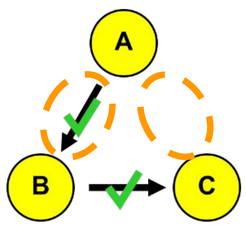
Explained Edge



People are more likely to infer reverse connectivity if they can visually see the co-(non)activations

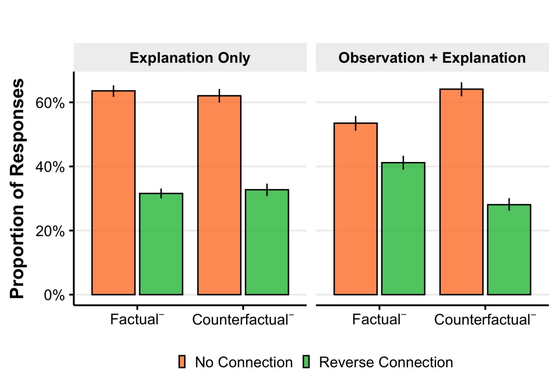
"If C didn't cause B, then perhaps B caused C."

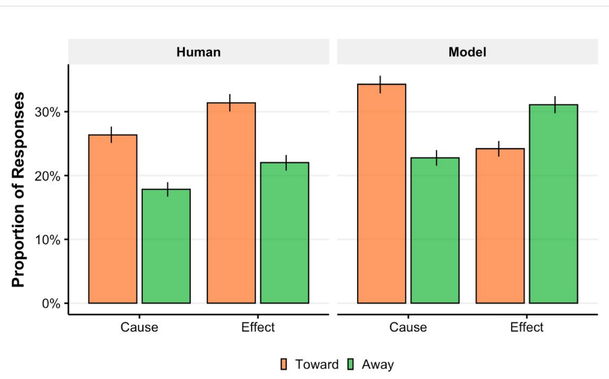
Unexplained Edges



People are more likely to infer the third component as a cause than as an effect

"If C didn't cause B, then perhaps A caused B."





Conclusion and Future Directions

- People benefit from explanations when making causal judgments, but deviate from optimality
- It is unclear how people integrate observations and explanations together
- There may be pragmatics involved in how people interpret and possibly communicate explanations
- Future studies could add temporal information to observations and interventions on the system